



MAX-PLANCK-GESELLSCHAFT

**HOCHSCHULE  
MITTWEIDA**  
UNIVERSITY OF  
APPLIED SCIENCES



---

# Bachelor Thesis

---

Misster

**Ron Hübler**

## **Discovery of variation in primate selenoproteins**

Mittweida, 2012

Faculty MNI

---

## **Bachelor Thesis**

---

# **Discovery of variation in primate selenoproteins**

Author:  
**Misster**

**Hübler, Ron**

Direction of study:  
**Biotechnology/Bioinformatics**

Seminar-group:  
**BI09w1-B**

First corrector:  
**Prof. Dr. rer. nat. habil. Roebbe Wünschiers**

Second corrector:  
**Dr. Sergi Castellano**

Submission:  
**Mittweida, the 13.12.2012**

Defense/Note:

**Bibliographische Beschreibung:**

Hübler, Ron: Variation discovery in primate selenoproteins - 2012. –7, 30, 1 S. Mittweida, Hochschule Mittweida, Fakultät MNI, Bachelorarbeit, 2012

**German Title:**

Untersuchung der Varianz innerhalb von Selenoproteinen von Primaten

**Short description:**

From a dataset consisting of 20 human, 20 chimpanzee, 20 bonobo individuals we created a catalogue of SNPs at 52 target genes. These target genes are either coding for selenoproteins, the cysteine containing analogue and orthologue genes, or are part of the synthesis machinery. We used GATKv.1.6 to call for SNPs at these specific genetic sites. After strict filtering the data was split gene wise and uploaded for SelenoDB 2.0

## **Directory of Content**

Directory of Figures .....	III
Directory of Tables.....	IV
Directory of Shortcuts .....	V
1. Introduction .....	1
1.1 Selenocysteine .....	1
1.2 SNP .....	2
1.3 Hominidae .....	4
1.4 Next Generation Sequencing (NGS) and Exome-capturing .....	4
2. Goal .....	7
3. Methods .....	8
3.1 Workflow .....	8
3.1.1 Overlook .....	8
3.1.2 Initial SNP-call.....	9
3.1.3 Annotation of the VCF-files .....	9
3.1.4 Filtering.....	9
3.1.5 Data-quality.....	12
3.1.6 Upload to the database .....	12
3.2 Tools .....	13
3.2.1 BAT.....	13
3.2.2 BED-Tools .....	13
3.2.3 Galaxy .....	13
3.2.4 Genome Analysis Toolkit (GATK) .....	14
3.2.5 GST .....	14
3.2.6 IGV .....	14
3.2.5 Reportmaker.....	15
3.3 File-extensions 3.3.1 SAM (Sequence Alignment/Map-format) .....	15

3.3.2 VCF (Variant Call Format).....	15
3.3.3 GFF/GTF (General Feature Format).....	15
3.3.4 BED (Browser extensible Data).....	15
4. Results .....	16
4.1 Results of filtering .....	16
4.2 Results of the examination of data quality .....	18
5. Discussion .....	22
5.1 Number of SNPs .....	22
5.2 High coverage in control-regions .....	22
5.3 Low coverage in some genes.....	23
6. Perspectives .....	27
7. Summary .....	28
8. Zusammenfassung.....	29
References .....	V
Appendix .....	VIII
Statement of authorship(Selbstständigkeitserklärung).....	IX

## **Directory of Figures**

Figure 1 Phylogenetic relations between Primates	4
Figure 2 Working with NGS- data	5
Figure 3 workflow of <i>SureSelect</i> starting with a prepared NGS-library	6
Figure 4 Workflow	8
Figure 5 Removal of indels	10
Figure 6 Steps in the coverage-cutoff	11
Figure 7 Average coverage at CDS regions	18
Figure 8 Average coverage on target regions per individual	19
Figure 9 Average coverage per gene for CDS part I	19
Figure 10 Average coverage per gene for CDS part II	20
Figure 11 Average coverage per gene for all features	20
Figure 12 Visualization of the <i>IGV</i> output at a control-region	23
Figure 13 <i>IGV</i> -output for <i>SelO</i>	24
Figure 14 Distribution of GC-content in all regions divided	25
Figure 15 Density plot for CDS regions	26

## **Directory of Tables**

Table 1 Number of SNPs in human	16
Table 2 Number of SNPs in bonobo	17
Table 3 Number of SNPs in chimpanzee	17
Table 4 Number of SNPs per gene	27

**Directory of Shortcuts**

BED	Browser extensible Data
DOI/DI	Iodothyne Diodinase
GATK	Genome Analysis Toolkit
GWAS	genome wide association study
RFLP	Restriction Fragment Length Polymorphisms
NGS	Next Generation Sequencing
SAM	Sequence Alignment/Mapping
SBP2	Selenocysteine binding Protein 2
SECIS	Selenocysteine-insertion-sequence
Sec- tRNA <sup>sec</sup>	Selenocysteyl-tRNA
Sel/SeP	Selenoprotein
VCF	Variant Call Format



## **1. Introduction**

The genetics department of the Max Planck Institute for evolutionary Anthropology in Leipzig provided a dataset consisting of the exomes of 20 human (Yoruba), 20 chimpanzee (Congo) and 20 bonobo (Congo) individuals. Our goal was to produce a catalog of single nucleotide polymorphisms (SNP) for 52 genes that have a relation to selenocysteine. This data will be uploaded to SelenoDB, a SQL-based database for genetic information on selenoproteins used and updated by Dr. Sergi Castellano's research group.

### **1.1 Selenocysteine**

Selenocysteine is the 21st amino acid in the genetic code and is encoded by reinterpreting the UGA-stop-codon [Stadtman, 1998]. Selenocysteine is a cysteine-analogue with two major differences. Selenocysteine uses a selenium-atom instead of sulfur and it relies on its own synthesis machinery, whilst cysteine uses the canonical synthesis pathways [Castellano et al, 2009]. The selenocysteine synthesis machinery uses tRNAs that are complementary to the UGA stop-codon. These tRNAs are amino-acylated with serine. These seryl-tRNA is converted to selenocysteyl-tRNA. The Sec- tRNA<sup>sec</sup> binds to a set specific elongation factors. A necessary factor is a complex formed by the SECIS (Selenocysteine-insertion-sequence)-element, which is downstream the selenogene, and eFFSec (eukaryotic selenocysteine-specific elongation factor) and SBP2 (Selenocysteine binding Protein). This complex is also required to reinterpret the in-sequence UGA-stop-codons [Gonzales-Flores et al, 2012] for selenocysteine.

The tRNA is delivered to the ribosome after binding to the elongation factors. Selenocysteine is common in vertebrates and results in dependence of the trace-element selenium [Castellano et al, 2009]. Selenocysteine is more redox active when being compared to cysteine analogue enzymes [Pamee, 2007]. The main function of the selenoproteome seems to be the reduction of oxidative stress at cellular level [Gonzales-Flores, et al, 2012]. However other selenoproteins have key positions in the metabolism for example Iodothine Diodinase (DI/DIO), which acts the thyroxine pathway [Marma, 2012]. Other selenoproteins like Selw1 [Li, 2012] and SelP [Pitts et al, 2012] seem to possess a function in the brain. Their gene-knock-out leaves lab-mice with learning dysfunctions. The human organism contains 25 selenoproteins, six cysteine-paralogues and four cysteine orthologues [Castellano, 2009].

## 1.2 SNP

DNA consists of the four bases Adenine, Guanine, that are called purine bases and Thymine, Cytosine, called pyrimidine bases, deoxyribose and phosphate. A sophisticated DNA-replication system is used by organisms. The DNA double helix opens up and the existing DNA-strands are used as a synthesis template. However mistakes happen and sometimes a wrong base is inserted or the DNA repair machinery replaces a correct position [Freeman, 2001]. The DNA-repair-machinery is orienting by special factors within the DNA-sequence as well as the properties of the DNA-loop structure [Eckstein, 1998]. Defects in the repair system will lead to the accumulation of mutations as well to some diseases, for example the Werner-syndrome (premature aging) [Eckstein, 1998]. Transition is the name given to the event when a purine base is replaced by another purine base. A transversion is a change between a purine and pyrimidine base. Transversions can be easier detected by the DNA repair system, because they disrupt the DNA-helix [Freeman, 2001]. The average error rate of DNA-replication lies between  $10^{-9}$  and  $10^{-10}$ , depending on the local gene environment [Jackson, 1996]. A SNP is a position in the genome where at least two bases exist within the population, the more common name given is point mutation, but it implies a negative fitness-effect.

At the beginning of genetics, before the verge of modern sequencers the only way to get polymorphism data was to run a gel electrophoresis with the target genes [Hartl, Clark, 2007]. This approach was possible because the proteins in a electrophoresis is very dependent on the length of the gene as well to the sequence, since it influences the electromagnetic properties of a amino-acid-sequence [Hartl, Clark, 2007], however every information on synonymous sequence differences will be unavailable [Hartl, Clark, 2007]. Some of the first Polymorphisms were discovered with the Southern Blot and were called RFLP (Restriction Fragment Length Polymorphisms) [Kwok1 et al, 2003]. Other methods that could be used include RNA cleavage of mismatched DNA; RNA that does not bind perfectly with DNA will be denaturized, or the Denaturing Gradient Gel Electrophoresis. The Denaturing Gradient Gel Electrophoresis is a method that utilizes the fact that even slightly mismatched DNA will denature earlier and will therefore take more time to move through the gel [Kwok1 et al, 2003]. The next step in developing reliable polymorphism-data was the gene shotgun, which ultimately lead to the next generation sequencing methods. The number of polymorphisms is systematically too low in the gel-electrophoresis, because small changes may not be detected unless the pH-level is varied [Hartl, Clark, 2007]. The same approach can also be used at the DNA- level. Restriction enzymes can shatter the DNA into specific pieces.

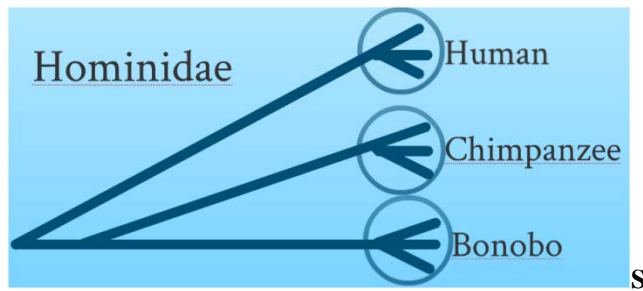
The human genome approximately contains between three and six million SNPs scattered across all regions [Kwok1 et al, 2003].

There is approximately one SNP per 1000 bp (base pairs) [Altshuler et al, 2008]. 90 % of human in-species variance is probably [Neumann, 2000] due to SNPs. Species-specific variance can be non-synonymous, which means the amino-acid sequence is changed or synonymous. Synonymous mutations do not change the amino-acid sequence. Regardless of this distinction most SNPs are recessive and remain neutral to the forces of selection [Altshuler et al, 2008]. However changing environmental conditions can trigger a recessive allele to become dominant. SNPs can be understood as a sort of unused genetic potential that will be realized when the right conditions are established [Freeman, 2001]. However the vast majority of the human genome has regulatory or unknown functions, SNPs that fall into these regions are much harder to evaluate [Wiler, 2003]. SNPs that influence the binding-efficiency can drastically influence the protein synthesis-efficiency of certain genes. In a broader context SNPs can be seen as a form of in-species variance that can, if it accumulates, result in a speciation event, thus leading to two separate species. In this context SNP are future inter-species variance [Barnes, 2003]. Regarding all realms vertebrates are the least polymorphic when compared to plants and invertebrates [Hartl, Clark, 2007].

If SNPs act at protein sites, they can have a fitness effect; it is therefore possible to establish SNP-catalogs of advantageous and disadvantageous mutations [Neumann, 2000]. Many disease associated loci fall into non-coding regions, which was shown by a genome wide association study (GWAS), evaluating DNase I hyperactivity sites that are connected to disease carriers [Maurano, 2012]. A GWAS is a study that aims to correlate genetic regions with known diseases. DNase is an enzyme that degrades DNA-strands into smaller entities. However the vast majority of the genome consists of non-coding regions and the DNA-repair machinery seems to be biased in a way that it favors coding regions [Jackson, 1996]. The higher number of mutations that exist in non-coding regions makes it therefore more likely to find deleterious ones. Some genes and regulatory parts of the DNA will only be activated at a specific stage of live or in a specific tissue. It is therefore necessary to address this. Maurono could show that 88.1 % of SNP covered in GWAS, fall into regions that are active at the fetal stage [Maurano, 2012]. Today SNPs can be found for example in dbSNP, a NCBI database. However due to the fact that many studies use exome data, SNPs in these regions can be somewhat overrepresented [Barnes, 2003]. Within the 3 billion base pair genome of a human only between 20.000 and 40.000 coding genes exist, which account for 3 % of the human genome.

### 1.3 Hominidae

The phylogenetic relationship between human, bonobo and chimpanzee is especially close.



**Figure 1 Phylogenetic relations between Primates**

Human, chimpanzee and bonobo belong to the group of Hominidae. The group of hominidae has a common ancestor. That ancestor later divided into two subspecies one leading to humans the other to chimpanzee and bonobo that split into separate species after the separation with humans [Springer et al, 2012]. However divergence exists also between the individuals of the same species.

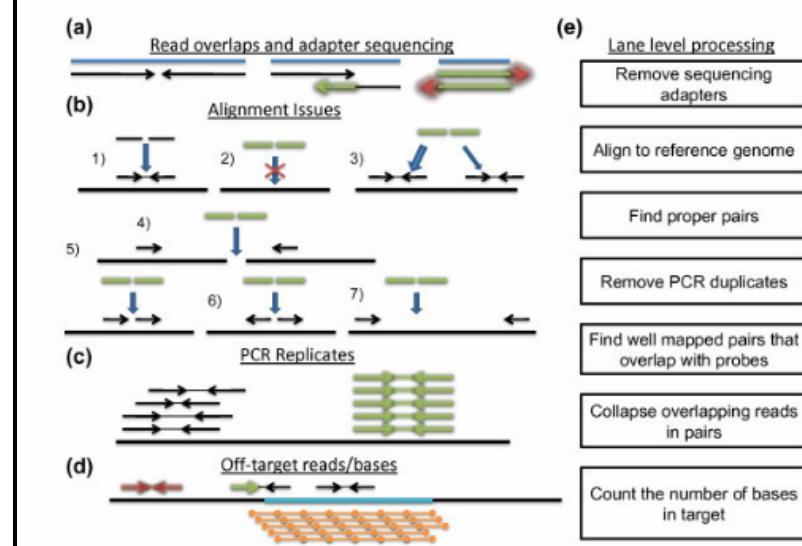
However the lifestyle of our closest relatives is still not completely understood. Many studies are conducted with zoo animals and have yet to be confirmed with studies of wildlife bonobos and chimpanzees [Tulke et al, 2008]. One of the most intriguing facts is the existence of cultural heritage and divergence in chimpanzees and bonobos [Tulke et al, 2008] and it raises the question if these different lifestyles manifest in in-species variance. Bonobo and chimpanzee belong to the few primates beside human that organize hunts and eat other vertebrates [Richard et al, 1985]. Additionally SNP data can be used to salvage information on the demographic structure and the relationships of wildlife primate groups. This information could help to stabilize the shrinking population sizes. Chimpanzees from central Africa are circa twice as divergent as humans [Parla, 2012]. Keeping in mind the close relationship between the hominidae [Prüfer et al, 2012] the divergence within the bonobo population will probably be close to either human or chimpanzee.

### 1.4 Next Generation Sequencing (NGS) and Exome-capturing

Sequencing technology has improved rapidly since the Human Genome Project. Sequencers are working with high levels of parallelization and are thus much faster and reliable. However the amount of data produced is huge. The 1000 Genome Project Pilot alone includes 5 terrabases of data [Mc Kenna et al, 2010]. Other Projects utilizing NGS-data is the Cancer Genome Atlas [E. Banks]. Due to the very small size of the reads (15bp-100 bp) which are mapped to a reference genome, NGS-data has relatively high error rates [Pattnaik, 2012]. Because of the large amount of data produced by NGS-sequencers it is somewhat preferred to

only establish Exome-datasets to minimize the amount of information [Ng et al, 2009] to a level where it can be easier handled and analyzed [Pattnaik, 2012].

### Description on working with NGS-sequencer-output



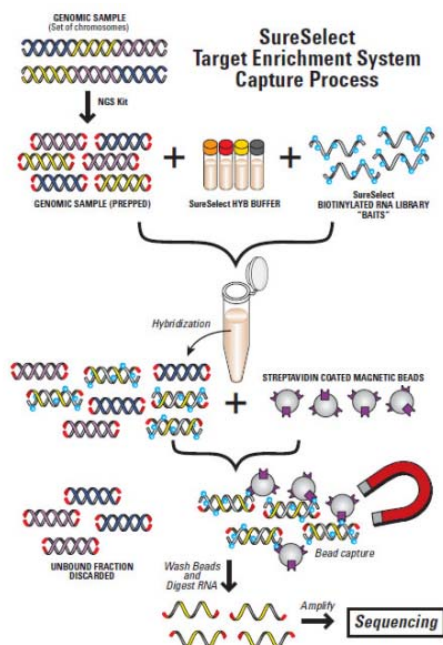
**Figure 2 Working with NGS- data [Albert et al, 20007]**

A summary of how to work with NGS-data. The first step in producing usable data out of NGS-sequencer output is to read and find overlapping positions and to remove the adapters. The second step is very crucial in producing reliable data- the mapping of the reads from the sequencer to a reference genome. Miss-mapped reads will lead to false positive signals in the analysis of the data. To map reads rightly it is helpful to identify regions that overlap with the gene probes used for the sequencing. The third step is to order the PCR replicates to the right mapped reads. The last step is to compute the number of reads that fall into target regions.

Exome- capturing became possible, because tools were developed that allowed the capture of target sequences directly and to bind them to microarrays [Albert et al, 2007]. The target sequences that are bound into dense microarrays allow parallelization and therefore NGS-technology [Albert et al, 2007]. It became a common approach to target exons to identify in-species variance at coding sites [Parla, et al, 2011]. There are certain commercial kits available to capture exomes. We used *SureSelect*, which was originally designed for human but was also used for primates in previous studies [Jin et al, 2012]. The greatest advantage of modern exome capturing is that 20 fold less DNA is needed as for whole-genome-sequencing [Parla et al, 2011]. The main disadvantage of exome-capturing however is that any variants present in structural non-coding regions will not be captured [Ng et al, 2009]. The data provided for this experiment originates from 20 humans from the Yoruba-population, 20 chimpanzees and 20 bonobos individuals. The bonobos and chimpanzees originate from

Congo (Central Africa). All exomes were mapped to “hg 19” the latest human reference genome.

Exome capturing is accomplished by using the properties of DNA-molecules, before all the H-bridge binding between the organic bases of complimentary DNA-strands.



**Figure 3 workflow of *SureSelect* starting with a prepared NGS-library**

[[http://www.genomics.agilent.com/files/Media/SS\\_Halo/Magnet584.jpg](http://www.genomics.agilent.com/files/Media/SS_Halo/Magnet584.jpg) (accessible 29.11.2012)]

Figure 3 describes the function of *SureSelect* the commercial exome-kit used to produce our dataset which is a relatively common used tool in exome capturing [Mc Donald et al, 2012]. The basic idea is to capture the target-DNA-regions in solution with specific RNA-baits that will bind the shattered DNA. These baits or probes are able to attach themselves to magnetic beads. The beads can easily be removed from the solution using a magnetic field. The captured regions will be attached to a microarray and sequenced.

## **2. Goal**

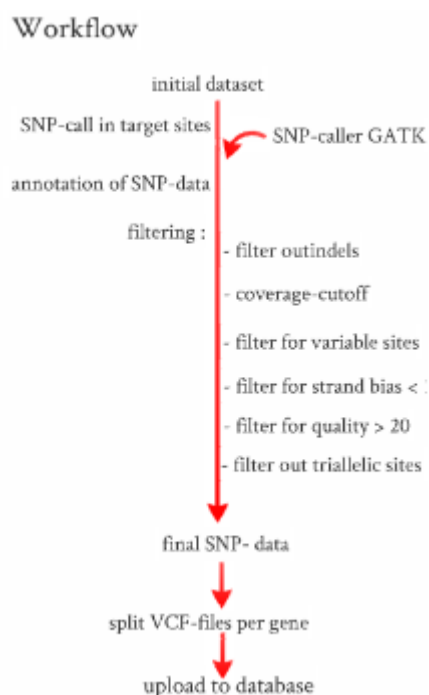
The goal of this endeavor is to produce a catalog of coding variation in selenoproteins genes and genes involved in selenium metabolism in humans, chimpanzees and bonobos. To this end we sample 20 humans, 20 chimpanzees and 20 bonobos and incorporate the variants into SelenoDB our database of selenoprotein gene annotation. All programs and data needed to meet this and was provided by the Max Planck Institute for evolutionary Anthropology's genetic department.

### 3. Methods

#### 3.1 Workflow

##### 3.1.1 Overlook

For our dataset, we had 20 BAM-files per individual in each species and one merged BAM-file per species. The Exome-capturing was already done and the files had already been filtered for mapping-quality. BAM-files are binary version of SAM-files, which are files that represent a sequence alignment on NGS-data [Li et al, 2009].



**Figure 4 Workflow**

A visualization of the steps we conducted in order to produce the SNP-data

We start with the merged BAM-files per species as initial dataset (also see Fig. 4). These BAM-files include the sequenced exomes. As the first step we call for SNPs in our target regions for each BAM-file separately. The output in this step will be a VCF- file per species that contains information for each position in respect to the same position in the reference genome. A VCF-file represents information on variants per position.

In this step the files contain positions that are unchanged when being compared to the reference genome as well as all changes in the sequence (SNPs, insertions, deletions). Information on the quality per SNPs and the individuals they were present is also included. Next we will annotate the files, meaning that we will add information on the ancestral allele



into the VCF-files which was previously not included. In the last step we apply various filters to the data to ensure the quality of the SNP catalog. The final dataset will be three VCF- files that only contain vary from the reference genome or between the 20 sequenced individuals. These three VCF-files will as a last step be split gene wise to allow the upload into the database.

### 3.1.2 Initial SNP-call

For the initial SNP call we used BAM-files that contained exome-data for all 20 individuals per species. This means per species we will be able to detect SNPs with a frequency of  $1/2n$  [Nei, 1987]. In our experiment the frequency is  $1/40$  or 0,025. *GATK* v1.6 (Genome Analysis Toolkit)[Mc Kenna et al, 2010, Depristo, 2011] was used as SNP-caller. The genomic regions that were targeted by *GATK* [Depristo, 211] consisted of a list of 50 loci that have a connection to selenocysteine or the selenocysteine-synthesis machinery. A full list with can be found in the Appendix or table 4. Additionally to the exons we called the control-regions and 200 bp (base pairs) of the neighboring introns. All sequences have been mapped to “hg 19” the latest human public reference-genome [Parla et al, 2011].

### 3.1.3 Annotation of the VCF-files

For the database it necessary that the VCF files contain information on the ancestral allele. To provide this information an institute internal python script- the *addAncestralAlleleFromAlignments.py* was used. The information is located in the header section of the VCF files. A copy of the headers was saved in an extra directory for later usage. Some of the steps necessary in the filtering will remove the header from the files, in which case the header has to be replaced manually after the filtering is done. The header will be removed each time we use *intersectBed* [Quinlan, 2012] to remove positions from the VCF-files.

### 3.1.4 Filtering

#### 3.1.4.1 Summary of filtering

Due to the properties of NGS-datasets (very short reads that can easily be miss-mapped), it is hard to separate real in-species variance from false positives [Ng et al,2009] created in the sequencing of the data [Depristo, et al, 2011 ]. Taking this into regard a set of filters was applied to the dataset, to remove not trustworthy positions.

The steps for filtering were (also see figure 4):

Step 1 Removal of insertions and deletions (indels)

Step 2 Coverage-cutoff (removal of low coverage positions)

Step 3 Extraction of variable sites

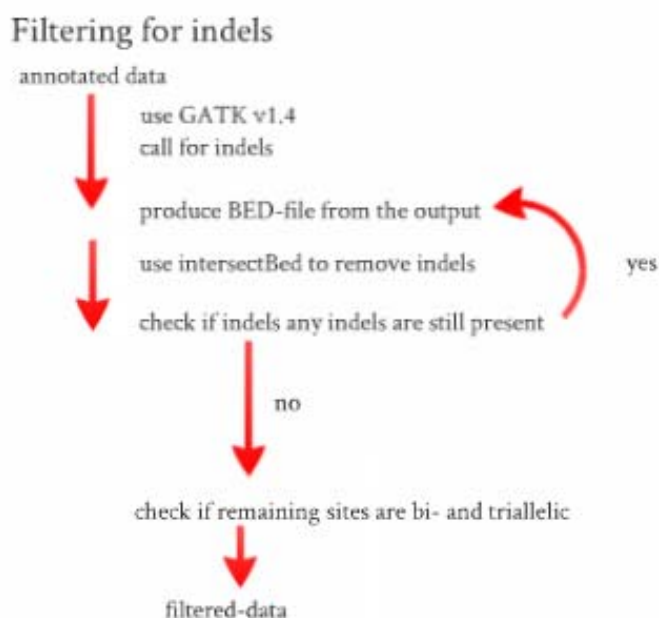
Step 4 Extraction of sites with quality scores above 20

Step 5 Extraction of sites with a Strand Bias below 10

Step 6 Removal of triallelic sites

### 3.1.4.2 Filtering for indels

The first step in the filtering is the exclusion of insertions and deletions (indel-positions) from the VCF-files per species.



**Figure 5 Removal of indels**

We start filtering with the annotated VCF-files. To find indel-positions, we call for them at the BAM-files and generate a BED-file with the positions to remove. After the indels have been excluded. It is necessary to check if all indels have been removed all from the VCF-files.

It is necessary to remove indels, especially if your sequencer produces short reads. In which case it is likely that indels are a product of miss-mapping [Ng et al, 2009].

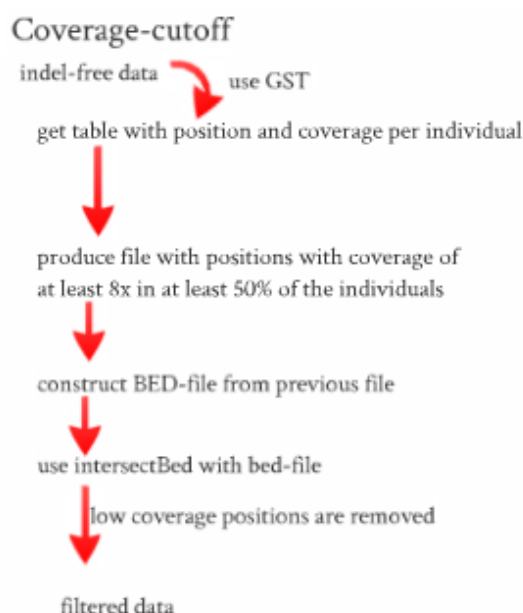
The filter is applied to the annotated VCF-files per species. The basic idea behind this step is to use *GATK* [Mc Kenna et al, 2010] and call for indels. The output will contain every indel-position. With this information it is now possible to construct a BED-file containing positions with indels. These positions can simply be excluded from the annotated VCF-files.

To remove the indels we used *GATK* v1.4 [Mc Kenna et al, 2010] and set it to detect indels (by setting *-glm* to *INDEL* and output mode to emit Variants only). This is a necessary step,

because SNPs in indels cannot be trusted due to the properties of NGS-data [Depristo, 2011]. The change in the *GATK* version was necessary because *GATK* v1.6 [Mc Kenna et al, 2010] calls triallelic sites at both SNPs and indels. At is therefore difficult to differentiate one from another. The output is a VCF-file that only contains insertions and deletions. We used this VCF-file to create a BED-file with all positions to remove. With the `-v` parameter in *intersectBed* [Quinlan, 2012] these positions could be removed from the annotated VCF-files. However, after checking, we could confirm that some indels remained in the files. So this step had to be repeated. All remaining sites that had more than one position in either allele column were triallelic, we so concluded that all indels had been successfully removed.

### 3.1.4.3 Coverage-cutoff

The next step in the filtering was the coverage cutoff.



**Figure 6** Steps in the coverage-cutoff

We start filtering with the indel-free dataset. In order to get low coverage positions we use *GST* to produce a table with coverage per individual. We construct a BED-file with low coverage positions and exclude these.

With the coverage-cutoff we try to remove positions from the indel-free VCF-files that have low coverage.

Coverage is a measurement on how often a specific position is present in the dataset. Low coverage means that there were only few reads in the sequencing, making it probable that these positions are miss-mapped. In the first step we used *GST* version 0.2 to produces a table per species with the locus, position, the average coverage and the coverage per individual. As input we provided one BAM-file per individual for all species.

From the *GST*-table a BED-file was created with all positions that had a coverage value of at least eight in at least ten individuals per species. We used *intersectBed* [Quinlan, 2012] again to keep only the positions contained in the BED-files for each species.

#### 3.1.4.4 Final filtering

We then extracted the variable positions from the VCF-files created in the previous step. Variable positions are all positions that do not contain a dot in the alt-column of the VCF-files. These positions are variable in the sense that they differ from the reference genome.

In the next step of the filtering all positions that have a strand bias-score above ten and a quality-score below 20 were excluded. These scores are already included in the VCF-files. In the last step we only kept positions that were not triallelic.

Triallelic is position in the VCF-file where more than one entry exists in the alt-column, which is most likely a result of miss-matching. It is very unlikely that one position should have three SNPs that are common enough to be detected in our dataset.

#### 3.1.5 Data-quality

To examine the quality of our SNP-data we used *BAT* and *Reportmaker* to get statistics on the quality of the dataset. Additionally we used *IGV* to plot reads at some genes from the BAM-files. To provide the input needed by *BAT* we intersected the BAM-files per individual which were also used for the coverage-cutoff with our target-region GTF-file [Quinlan, 2012]. The resulting output was piped into an extra directory and later fed to *BAT*. We used *BAT* twice once with the original GTF-file, that contains the CDS (coding sequence), 200bp of the introns, the control-regions, the 5'- and 3'- UTR (untranslated region), upstream and downstream regions as well as the tRNA-regions. The second time we only analyzed at the CDS. *Reportmaker* used the *BAT* output to produce graphs that gave some information on the distribution of coverage within the dataset. Because the coverage in some genes was low (around three and five when the average of coverage is supposed to be around 20x), *IGV* was used to see where exactly the reads fall in the BAM-files [Robinson et al, 2011].

#### 3.1.6 Upload to the database

In order to implement the data into the SelenoDB it was necessary to separate the filtered VCF-files in gene wise manner per species. We used *intersectBed* [Quinlan, 2012] to split the filtered VCF-files along the regions that were specified in the BED-file we used at the initial SNP-call. However we did not split for the control-regions, which were also included in the BED-file. They are insignificant for the database. The SelenoDB is a SQL-based database for genetic information on selenoproteins. The database consists of specific data-layers that must

also be addressed while presenting data. To import the data it is necessary to present a hierarchical directory structure, for example “Species/Family/Subfamily”. Part of this information can be acquired from the headers of the filtered VCF-files. In the next step data-layer files must be added to the annotation directory. At first the author must be specified. Next information on the species and its taxonomy must be presented. The third-layer addresses the population. It must contain following information name, continent, country, genus, and species. Following this approach the next layers are for the individual’s data, the family and the sequencing and analysis technology, the sequence and the genetic features. Genetic features are for example exon-intron borders, the promoter and transcript sites. Two layers have to be created to describe the Sequence and Features in further detail and the last layer is for external references. The whole directory needs to be compressed and send to the database, where all the information specified will be expressed.

## 3.2 Tools

### 3.2.1 BAT

*BAT* is a house-internal Java-program written by Juan Ramón Meneu Hernández. It can be used to either call SNPs directly or to produce a table that contains the coverage at the target regions. For this step it is necessary to present a directory with BAM-files per individual as input.

### 3.2.2 BED-Tools

BED-Tools are a software-suite that can be used to examine SAM-files, GTF-files and BED-files [Quinlan, 2012]. We used *intersectBed* to extract overlapping features between BED-files, we made during the filtering and our VCF- files. During the examination of our data quality we also used it on some of the BAM-files, from the initial SNP-call. In the examination of the quality of our dataset we also resorted to *coverageBed*. This function can be used to look at the coverage per position.

### 3.2.3 Galaxy

*Galaxy* is a cloud-based online platform used to analyze biological and genomic datasets [Giardine et al, 2005]. The platform is designed in a fashion to make it intuitively useable for scientist with no experience in programming at all [Goecks et al, 2010]. The preprogrammed tools work like a customizable pipeline leading to a higher reproducibility [Blankenberg et al, 2010].

### 3.2.4 Genome Analysis Toolkit (GATK)

For this Project the versions 1.4 and 1.6 of *GATK* were used. *GATK* is a platform independent Java toolkit provided by the Broad Insitiute to deal with NGS-datasets. *GATK* is especially efficient in dealing with large datasets due to the MapReduce-approach, which divides the computation into two separate steps, the data management and the analysis [Mc Kenna et Al. 2010]. It uses a map and a reduce function. The map function divides the computation into walkers and traversals. [Mc Kenna et al, 2010]. Traversals prepare and divide the data into shards (some multi kilobase long data-fragments) and walkers that analyze data for the map and reduce function. The map function breaks the data into small independent parts and analyzes it and the reduce function maps the created output to the final result. The *GATK* core-module can be upgraded with additional tools to equip it for specific tasks. For the input BAM and SAM files are encouraged however most sequencer output can be converted into these formats by *GATK* [Depristo, 2011]. SNP-detection and genotype function work well with the map reduce function, because they both require analyzing each locus separately [Mc Kenna et al, 2010].

In a benchmark performed by Suretansu Pattnaik *GATK* showed to be the most accurate toolkit in variance calling but also the slowest [Pattnaik, 2012].

*GATK* can be used to call Variants. The figure shows the standard Variant call. You can start after the NGS Data is processed and with this data it is possible to call either SNPs, indels or SVs(Structural Variants), that will undergo a program internal analysis, to evaluate their quality [Depristo, 2011].

### 3.2.5 GST

*GST* is another house-internal Java-script written by Juan Ramón Meneu Hernández, which can be used to produce a table containing the (alignment) quality per individual. We used *GST.0.2.jar* for our analysis.

### 3.2.6 IGV

*IGV* is a scalable Genome-viewer that can be used to visualize NGS-data to get an overall expression for the file [Robinson et al, 2011]. Additionally it can be used to share files with other scientists.

### 3.2.5 Reportmaker

*Reportmaker* is a Perl-script from Frédéric Romagné who works at the MPI EVA (Max Planck Institute for Evolutionary Anthropology in Leipzig) it can be used to summarize the BAT-output graphically.

## 3.3 File-extensions

### 3.3.1 SAM (Sequence Alignment/Map-format)

The SAM -format contains aligned and mapped NGS-data. BAM is the binary version of a SAM-file and more compact. We use BAM-files to present the sequence reads per individual and per species. SAM-files are divided into a Header section and a sequence section [Li et al, 2009]. The header contains overall information the sequence section information per position in the sequence.

### 3.3.2 VCF (Variant Call Format)

VCF is a file format that is somewhat derived from BAM-format and is used to store information on genetic variation. We use the VCF-format to store our SNP-data. Like BAM/SAM-files VCF-files are also divided into a Header-section for general information and a sequence section for more specified information per position [Danecek, 2009].

### 3.3.3 GFF/GTF (General Feature Format)

The GFF-format is used to specify genetic information [Durbin, 2000], it is more restrictive as the BED-format allows however a clearer definition of genetic regions. We use the GFF-format to present our target regions to for *BAT*.

### 3.3.4 BED (Browser extensible Data)

The BED-format like the GFF-format can be used to specify genetic regions [Quinlan, 2010]. However in opposition to the GFF is the BED much less restrictive. Here we normally use it to specify positions for *intersectBed* [Quinlan, 2012] during the filtering.

## 4. Results

### 4.1 Results of filtering

At first we present a summary of the number of SNPs and sites contained in the VCF files after the SNP-call and after each step of filtering. The “Number of Sites” is the number of positions that are contained in the files. Everything that is not a header is counted, fixed differences, variable positions and positions that do not vary from the reference genome. For the “Number of SNPs” we count each position which differs within the 20 sequenced individuals. The first table presents that number of SNPs and variable positions in each step of the filtering for the human dataset.

**Table 1 Number of SNPs in human**

Human		
Step	Number of Sites	Number of SNPs
Initial SNP-call	2,111,133	1,456
Indel removal	2,110,011	1,350
Remaining indel removal	2,110,009	1,350
Coverage-cutoff	131,041	532
Filter for var. sites	n. a.	532
Filter for quality > 20	n. a.	497
Filter for Strand bias < 10	n. a.	494
Exclusion of triallelic sites	n. a.	494

After the initial SNP-call there are roughly 1,500 SNPs in the human dataset. After the last step of the filtering, the exclusion of triallelic sites, only 494 SNPs remain. Circa two thirds of the original SNPs were removed. The largest portion of SNPs is excluded during the coverage-cutoff, suggesting that these positions had low coverage and could rather have been false positives.



Table 2 is summary of the number of SNPs for the bonobo dataset.

**Table 2 Number of SNPs in bonobo**

Bonobo		
Step	Number of Sites	Number of SNP
Initial SNP-call	2,111,510	1,744
Indel removal	2,106,008	1,532
Remaining indel removal	2,106,008	1,532
Coverage-cutoff	140,131	559
Filter for var. sites	n. a.	559
Filter for quality > 20	n. a.	535
Filter for Strand bias < 10	n. a.	533
Exclusion of triallelic sites	n. a.	527

Bonobo starts with a slightly higher number of SNPs than human, 1,744 SNPs. But after applying all filters 527 remain. Also two thirds of the originally called SNPs were removed, most of them during the coverage-cutoff.

The number of SNPs during each step, for the chimpanzee data is summarized in table 3.

**Table 3 Number of SNPs in chimpanzee**

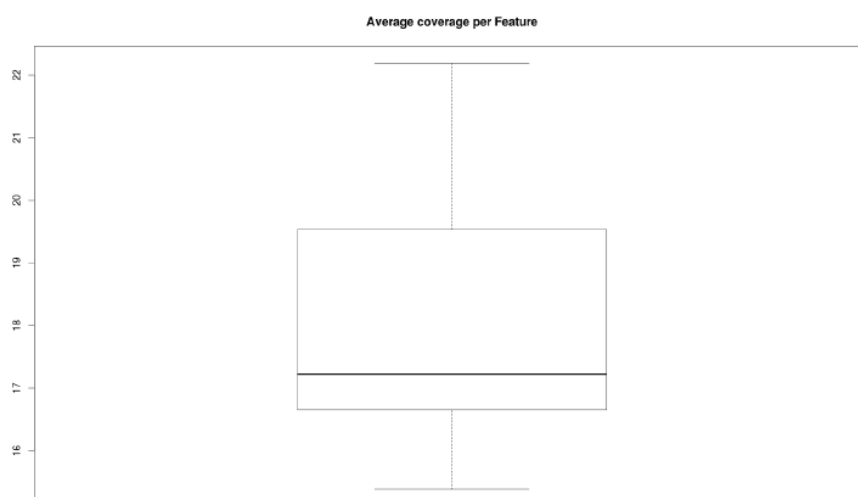
Chimpanzee		
Step	Number of Sites	Number of SNP
Initial SNP- call	2,111,495	3,549
Indel removal	2,105,764	3,180
Remaining indel removal	2,105,763	3,179
Coverage-cutoff	135,803	1,214
Filter for var. sites	n. a.	1,214
Filter for quality > 20	n. a.	1,159
Filter for Strand bias < 10	n. a.	1,153
Exclusion of triallelic sites	n. a.	1,143

In the initial VCF-file chimpanzee has more than twice as many as human 3,549. After the exclusion of triallelic sites 1,143 SNPs are left, roughly twice as many as in human.

The remaining SNPs account for 33.9 % of the initial SNP-data in human, 30.2 % in bonobo and 32.2 % in chimpanzees. During the coverage-cutoff 818 SNPs were excluded for human, 973 SNPs for bonobo and 1,965 SNPs for chimpanzee. In percentage, 56.2 % for human, 55.8 % for bonobo and 55.4 % for chimpanzee have been removed during the coverage-cutoff, which as mentioned makes it the strictest part of the filtering, in the sense that it posed as the highest threshold for a position to be included in the SNP-catalog.

## 4.2 Results of the examination of data quality

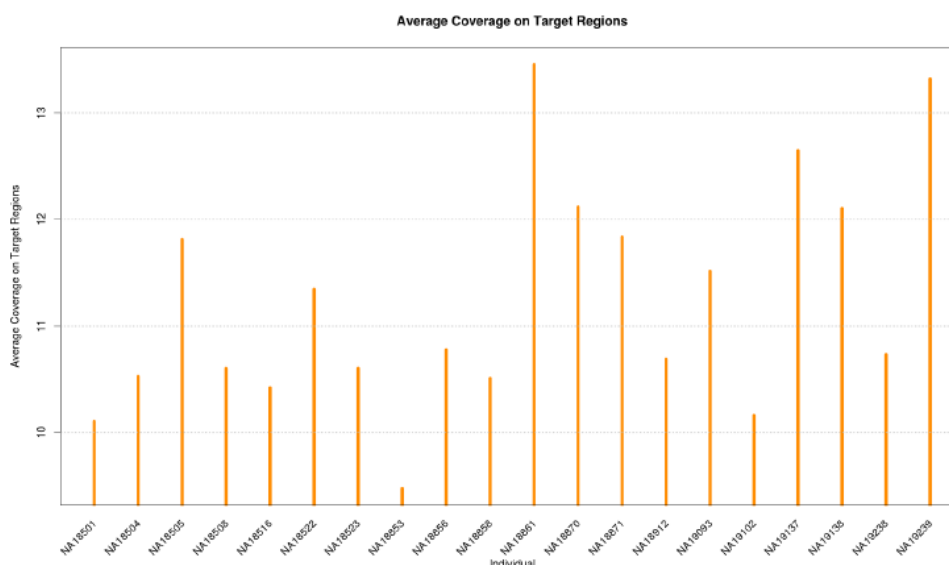
Because the coverage at the called sites was the factor with the highest influence during the filtering we chose to look at the average coverage of our called regions in more detail. To do so, we used *BAT* to get some statistics on the average coverage per feature; we visualized the output using *Reportmaker*.



**Figure 7 Average coverage at CDS regions**

The figure was produced by *Reportmaker* from the *BAT*-output

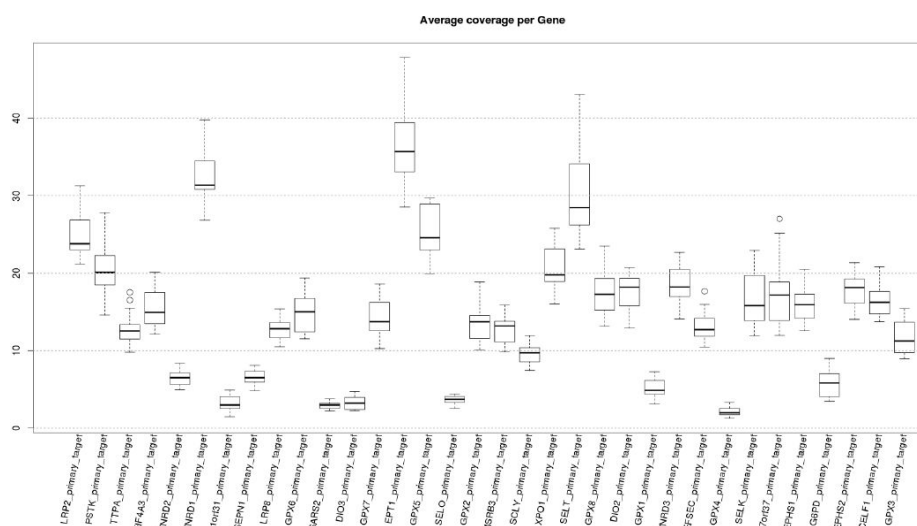
Figure 7 shows the coverage in CDS-regions. It has a statistical means at 17x and varies between 15x and 22x. Coverage of 20x was expected in the CDS regions. Judging from the plot (figure 7) the data quality is relatively close to that. Despite the fact, that the largest portion of SNPs was removed during the coverage-cutoff a coverage of 8x does not seem to be too strict.



**Figure 8 Average coverage on target regions per individual**

The plot was produced from the *BAT*-output for all-target-region.

The figure displays average coverage for all genetic features across the regions were SNPs were called per individual. Basically all individuals show an average coverage between nine and 13. The average coverage per individual is lower than for the CDS because exome-capturing kits are optimized to capture coding genes and not other genetic regions like UTR or control-regions, which still can be captured to some extent.

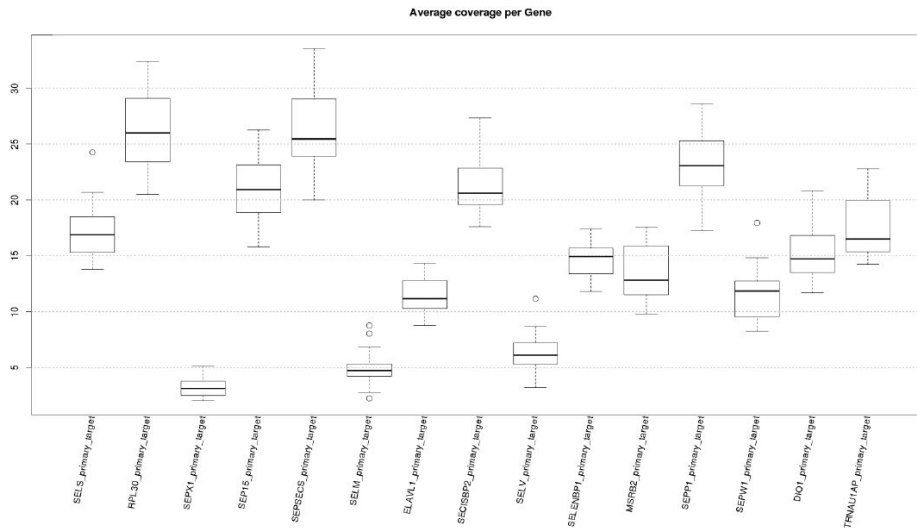


**Figure 9 Average coverage in all genes for CDS part I**

The plot gives the distribution of coverage in all CDS-regions that were targeted in the SNP-call.

Figure 9 is part of the *BAT*-output for CDS-regions, to get a clearer picture of the quality.

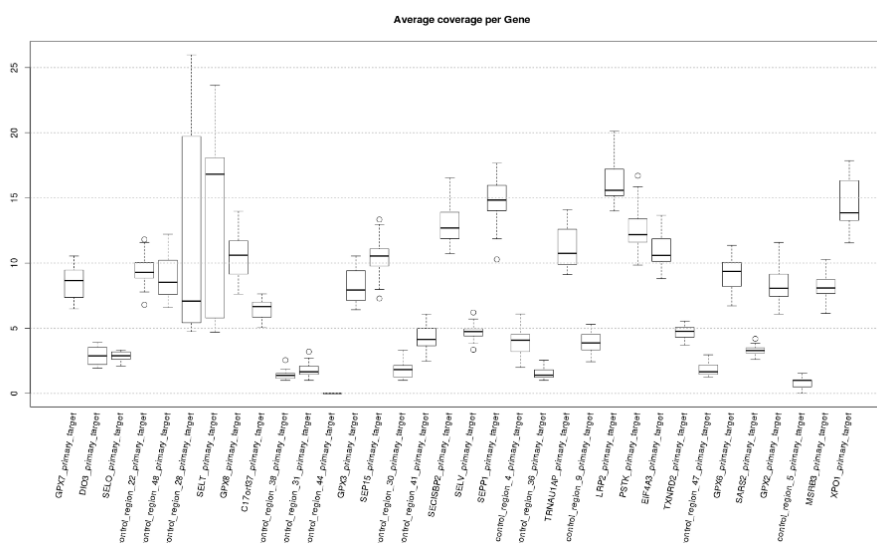
It can be seen that there are some very high coverage genes but also genes with a coverage around five. The genes with lowest coverage are C11orf31, SARS2, DIO3, and GPX4. The vast majority is between ten and twenty, which is expected taking figure 7 into account.



**Figure 10 Average coverage in all genes part for CDS II**

This plot complements figure 9 and displays the remaining target genes that were called but found no place in the previous plot.

The regions with the lowest coverage are SELPX1, SELM and SELV, which also somewhere around five. Most genes are also between ten and twenty and two genes (RPL\_30 and SEPSECS) show a higher coverage. It is visible that the divergence in coverage per individual is more homogenic in low coverage genes than in genes with higher average coverage. This suggests that whatever factor is causing the low coverage happens in all individuals.



**Figure 11 Average coverage per gene for all features**

Figure from the BAT-output for all features. It is one of three Plots that show the average coverage per gene

Figure 11 visualizes that most of the control-regions are lower covered than the coding genes. The average coverage in coding genes is lower in this figure when being compared to figure 9 and 10 because the UTR-regions and introns are also included in this plot. Commercial exome-capturing kits tend to be optimized in a way that allows the capture of CCDS (consensus coding sequences) in humans [Parla et al, 2011]. This explains why the coverage drops once non-CDS regions are included in the box-plot. However figure 11 also visualizes that some control-regions have coverage above ten, which is relatively high taking in regard that they weren't to be captured at all.

The coverage of the data was entirely what was we expected. Despite the fact that the average coverage in the CDS- regions lies around 16 and 17 some genes show drastically decreased coverage. First we used *IGV* [Robinson et al, 2011] to compute how the reads are distributed at target regions within the BAM-files. There seems to be tendency in a gene with low coverage to lose information at the edges of exons as well as the outer exons, mostly close to sequence end. There are also some control-regions that are relatively long uninterrupted sequences and seem to have higher coverage. It is also possible that some probes for them are included in the *SureSelect*-exome-kit.

All these figures were produced using the human dataset; however they do not vary from chimpanzee and bonobo. The genes that have low coverage and the genes that have high coverage are the same in each species, which excludes the occurrence of a species-exclusive phenomenon.

## **5. Discussion**

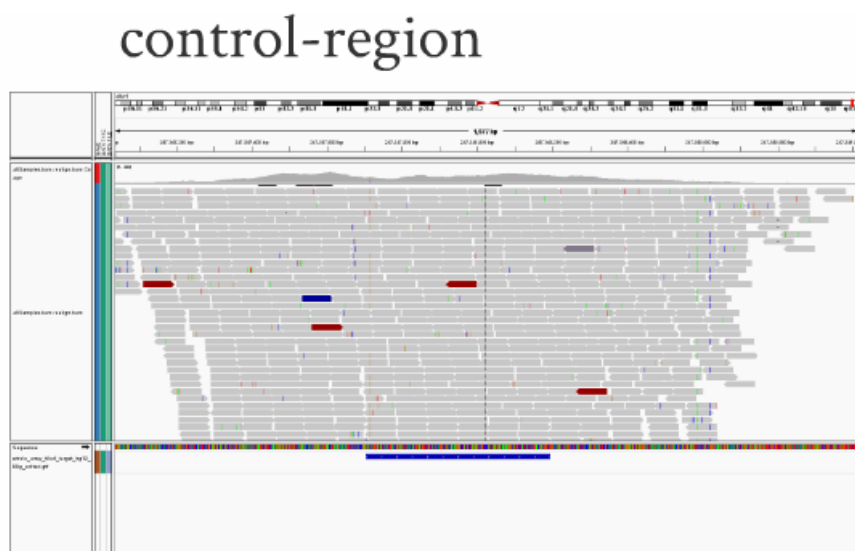
### **5.1 Number of SNPs**

In the initial VCF-files were 1,456 SNPs for human and 1,744 for bonobo and 3,549 for chimpanzee. After the filtering remain 494 SNPs in human, 527 in bonobo and 1,143 in chimpanzee.

There is roughly twice the amount of SNPs in the chimpanzee-dataset than in the human-dataset. That fact is consistent with previous studies that showed chimpanzee from Central Africa to be more divergent than humans [Hvilsom et al, 2011; Tarjei et al, 2005]. For bonobos the rate of polymorphisms is expected to fall close to humans and chimpanzees, due to the close phylogenic relation to both of them [Prüfer et al, 2012]. In our dataset we see the number SNPs to be slightly increased when being compared to humans, which is in the expected range.

### **5.2 High coverage in control-regions**

Control-regions were expected to show almost zero coverage, because commercial kits tend to target only CDS [Parla et al, 2011]. Control-regions tend to be small strands of DNA distanced from the genes they regulate [Brown, 2007]. Enhancers and Silencers are normally more upstream other factors can be closer to the promoter [Brown, 2007]. This is true for many of the control-regions, however some show relatively high coverage. Visualization in *IGV* [Robinson et al, 2011] showed that some control-regions are relatively large strands of DNA, which probably made them easy targets for the gene probes, thus explaining the high coverage.



**Figure 12 Visualization of the IGV output at a control-region**

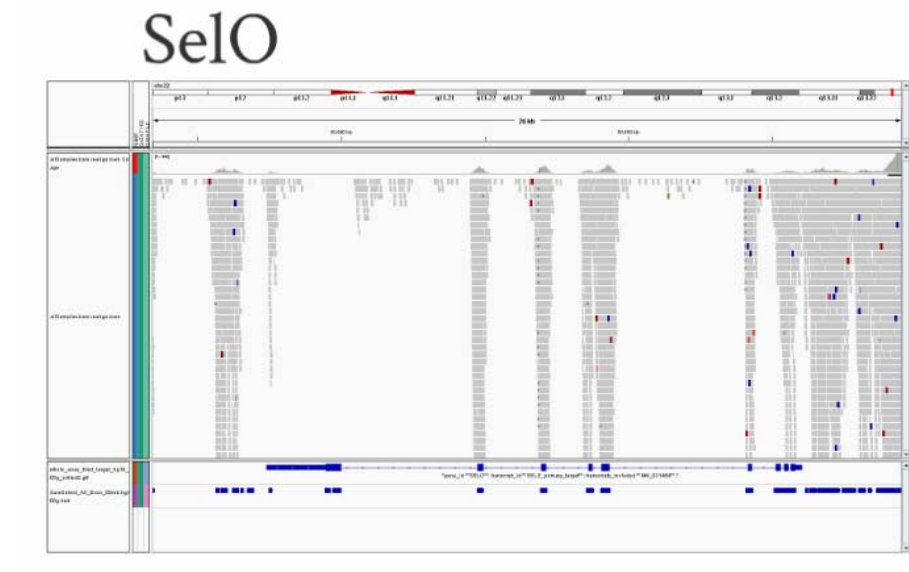
The blue line represents called region the grey fragmented lines displays the distribution of reads at this locus in the merged human BAM-file

Figure 12 shows a circa 2 Kb long control-region and it is visible that many hits fall around this region, resulting in a high coverage. This could either indicate that *SureSelect* contains gene-probes for some control-regions or be result of miss-mapping. It is also possible that the probes were able to target sites by chance without being optimized for it. However this is more of an intriguing fact than an issue that has to be faced for the dataset. SNPs at control-regions won't be moved to the database.

### 5.3 Low coverage in some genes

The next step was finding a solution why the coverage in some genes was relatively low, around 5x when the expected coverage is around 20x. The results of the filtering (table 1-3) show that the the greatest portion of SNP-positions is excluded during the coverage-cutoff in each species, roughly 50 % of our data per species have been excluded during the coverage-cutoff it is therefore necessary to address this. The 50 % of exclusion are consistent for each species. This indicates there is now species-specific factor causing this distortion. At first *IGV* [Robinson et al, 2011] was used to visualize the distribution of reads per exon. A tendency to have low coverage in outer exons and at the edges of some inner exons seemed to be visible.

There seemed to be tendency for small exons to have less coverage.



**Figure 13** IGV-output for SelO

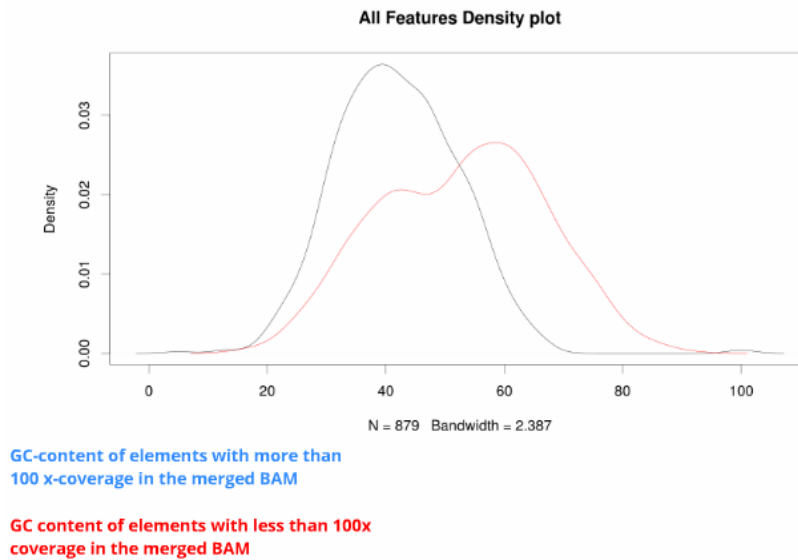
The first blue line represents the regions file we used for the *BAT* the second blue line display the gene-probes from *SureSelect*. The grey fragmented lines visualize the distribution of hits.

Coverage in outer exons and at some edges of inner exons tends to be lower (figure 13). The same pattern was observed in the visualization of other genes with low coverage. However without a mathematical way to correlate exon-length with average coverage per exon, these tendencies are not reliable and different factors must be taken into account.

In the next step target regions were being compared with the gene-probe-regions from *SureSelect*. If there would be a major divergence between these files we could hypothesize that the commercial exome-capture kit (*SureSelect*) is not designed in a fashion that allows the capture of selenoproteins. We used *intersectBed* [Quinlan, 2012] to compute the overlapping and non overlapping parts of these two files. Out of 1,655 sequence-elements that were used to call SNPs, roughly 200 did not overlap with the gene-probes-regions. But only four fell into CDS-regions and only one was in a gene known to have lower coverage. It can be conclude that the low coverage is not caused by an inability of the commercial kit *SureSelect* to capture selenoproteins. However if the properties of the local gene-environment in low coverage genes, somewhat differ from high-coverage genes they may be the influencing factor. The GC-content is known to have an influence on the efficiency of gene-capturing [Jin et al, 2012]. GC-content refers to the fraction of the bases Guanine and Cytosine in a part of DNA. Between Guanine and Cytosine three h-bridges established in double stranded DNA, which results in a more stable connection in opposition to Adenine and



Thymine which only establish two h-bridges [Brown, 2007]. A very low GC-content could therefore lead to an unstable connection between target-exon and gene-probe, a very high GC-content especially, in small exons could cause them to remain double-stranded and leave them uncatchable for gene-probes. *Galaxy* [Giardine et al, 2005] was used to compute the GC-content in per target-region and *coverageBed* [Quinlan, 2012] to get the average-coverage per region.

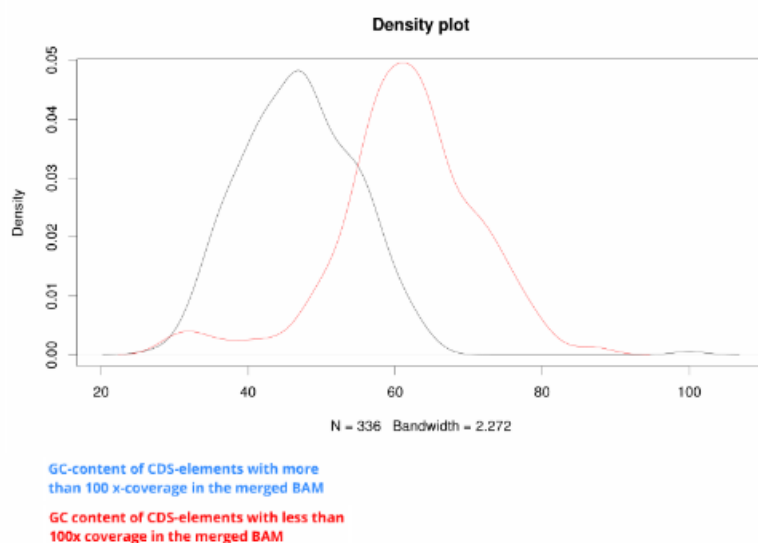


**Figure 14 Distribution of GC-content in all regions**

The plot shows the denistiy of GC-content. We seperated our data into two groups one group represents sequence-elements with above 100x coverage the other group the sequence-elements with lower than 100x coverage. We used the BAM-file that contains all individuals to compute the average coverage with *coverageBed* [Quinlan, 2012]. W expected a coverage of 20x per individual. In this merged BAM the average coverage should be around 400x(20\*20x). We computed the GC- content and the average coverage for the same regions we initially called SNPs at. We used R to compute the plot.

GC-content for low coverage-positions(red) is shifted to the right in opposition to the GC-content for high coverage positions (blue) which has a maximum at 40 and goes down to 20 and 40. Low coverage positions seem to have overall a higher GC-content the highest density lies at 60, there is a second maximum at 40. This suggests that the factor that influences our low coverage genes is the GC-content.

To be sure that the same pattern can be observed for CDS-regions, we also computed a similar plot only for CDS-regions.



**Figure 15 Density plot for CDS regions**

It was computed in same way as the previous figure. The only difference is that we only display CDS-regions in this plot to exclude the possibility that the pattern observed in figure 14 is a result of the other features.

When only the CDS-regions are examined the same shift can be observed for all features, but much clearer. The GC-content for low coverage positions (red) is shifted ca 20 % to the right when being compared to the high coverage positions (blue). The plot for high coverage CDS positions is very similar to the graph for all features. The maximum is still at 40 and we have lower density when we close in to 20 and 40. For low coverage positions we have the maximum at 60 and have lesser density closing in to 40 and 80. There is also a little peak at 30 GC-content suggesting that also very low GC-content influences the sequence quality. But at least in our data set not as much as high coverage.

The conclusion is that the genes that tend to have low coverage and should have been captured by the gene-probes. However the high GC-content in low coverage exons interfered with the ability of the correlating gene-probes to capture them. Because the same pattern of higher (or very low) GC-contents in all genes in all features can be seen in all features and the CDS, it is very likely that similar results would have been displayed in other genes with comparable GC-content.

After the splitting of the VCF-files a final summery was made containing the number of SNPs in each gene and species (Table 4). The number of variable positions is smaller in table 4 than in tables 1-3 because we only split the VCF-files for coding genes. Control-regions will not be uploaded to the database and were thus not present in these files. This reduced the number of variable positions. There are some genes that show an excess in variable positions, for example EIF4A3 and TXNRD1.

**Table 4 Number of variable positions per gene**

Gene	Human	Bonobo	Chimpanzee	Gene	Human	Bonobo	Chimpanzee
C11orf31	0	0	3	SCLY	14	6	22
C17orf37	2	1	4	SECISBP2	11	17	33
CELF1	8	8	14	SELENBP1	3	2	21
DIO1	7	2	6	SELK	4	6	5
DIO2	9	4	12	SELM	3	0	0
DIO3	0	0	0	SELO	0	6	1
EEFSEC	3	2	13	SELS	8	6	13
EIF4A3	16	70	114	SELT	10	6	19
ELAVL1	4	8	7	SELV	2	1	2
EPT1	18	8	21	Sep 15	7	7	9
G6PD	2	3	7	SEPHS1	4	3	12
GPX1	0	0	4	SEPHS2	2	4	2
GPX2	0	2	5	SEPN1	4	8	14
GPX3	3	4	6	SEPP1	8	9	14
GPX4	0	0	0	SEPSECS	9	10	34
GPX5	7	8	8	SEPW1	2	1	1
GPX6	8	5	5	SEPX1	2	1	2
GPX7	4	10	5	tRNA_17	0	0	0
GPX8	2	3	8	tRNA_19	0	0	0
LRP2	135	110	321	tRNA_22	0	2	2
LRP8	10	13	21	TRNAU1AP	8	10	30
MSRB2	3	5	12	TTPA	1	2	8
MSRB3	9	8	28	TXNRD1	27	9	68
PSTK	12	7	10	TXNRD2	6	9	21
RPL30	4	8	24	TXNRD3	17	11	26
SARS2	6	3	3	XPO1	19	24	36

There are low numbers of variable positions in GPX4, SELO, DIO3 and other genes that had low coverage. The number of SNPs in these genes might not be reprehensive and should be treated carefully when analyzing the dataset, because the true number of variation in these genes could be higher. Except for LRP2 an excess in the number of SNP-positions does not really seem to correlate between all species.

## **6. Perspectives**

The next logical step would be to analyze the dataset. First it would be possible to look at the distribution of SNPs per gene and if they correlate in each species. As it is visible in table 4 this does not seem to be the case judging from the raw numbers of SNPs. However a more qualified analysis not only measuring the amount of SNPs per gene but also setting them in a context to the divergence per species might reveal underlying correlations. It would also be plausible to gather information on the biological effects of each SNP. There is the possibility that there are SNPs present in some gene that could alter the behavior of the encoded protein.

In the near future there may be reliable and complete enough SNP data sets for all primates, to see how the genetic variance differs between these species and if there a proteins that are under positive selective pressure in some subpopulations. It would be intriguing to see if such cases if they exist correlate with local selenium levels and in which direction these proteins are pushed by selective pressure. For example there could be cysteine/selenocysteine exchanges or the other way around. To have information whether inspecies variance contains these differences could help to contribute to the picture of selenocysteine/cysteine interchangeability.

## **7. Summary**

We used *GATK* v 1.6 to call for SNPs in 52 genes in a dataset composed of 20 humans, 20 bonobos and 20 chimpanzee exomes that were captured with *SureSelect* and mapped to hg19. Each gene has a connection with selenocysteine. They were either genes of selenoproteins, cysteine-containing orthologues or paralogues to selenoproteins or part of the selenocysteine-synthesis machinery. Strict filters were applied to ensure the quality of the SNP-catalogue. Steps in the filtering were the removal of indels, the coverage- cutoff to exclude low coverage positions, the extraction of variable sites, filtering for quality above 20 and strand bias below 10, lastly the exclusion of triallelic sites. We had to exclude almost two thirds of the SNP-positions in each species. Most of these SNPs were removed during the coverage- cutoff circa 50 % of the initially called SNPs. A more thorough look at the average coverage per gene and feature revealed the low coverage in some genes and an unexpected high coverage in control-regions. The distribution of *SureSelect*'s gene probes showed that each gene should have been captured. However we were able to show that the low coverage in some exons is due to high and very low GC-contents. Therefore we found no proof for a systematic error in the *SureSelect* gene-probes that could result in an inability to catch selenoprotein-genes. Before uploading the data to SelenoDB, we split the SNP-files per species gene wise As a last step a summery on the number of SNPs per gene was computed.

## **8. Zusammenfassung**

Ziel war es einen SNP-Katalog von 52 Genen aus einem Datensatz bestehend aus den Exomen von 20 Menschen, 20 Schimpansen und 20 Bonobos zu erstellen. Die Exome wurden mit Hilfe des kommerziellen Kits *SureSelect* gewonnen und auf hg19 kartiert. Als Softwarepaket für den SNP-call wurde *GATKv1.6* benutzt. Alle Zielgene stehen in einer Verbindung mit Selenocysteine, sie sind entweder die Gene des Selenoproteoms, oder cystein-orthologen, -paralogen Gene, beziehungsweise Teil des Selenocystein-Synthese-Apparates. Um die Qualität des SNP-Katalogs zu sichern, wurde ein Satz strikter Filter benutzt. Zuerst haben wir Indels entfernt, danach wurde ein coverage-cutoff durchgeführt, später extrahierten wir die variablen Positionen, filterten für einen „quality-score“ größer 20 und einen „strand bias“ kleiner 10, als letzter Schritt wurden triallelische Positionen entfernt. Insgesamt wurden etwa zwei Drittel aller SNPs pro Spezies heraus gefiltert die meisten etwa 50 % der ursprünglichen SNPs während des „coverage-cutoff“. Bei genauerer Betrachtung der coverage (Deckung) in einigen Genen stellte sich eine sehr niedrige Deckung in einigen unserer Zielgenen heraus, allerdings wiesen einige Kontrollregionen eine sehr hohe coverage auf. Eine Korrelation mit niedriger coverage und hohen(bzw. sehr niedrigen) GC-content aufgezeigt werden. Es konnte kein systematischer Fehler entdeckt werdend er Selenoprotein spezifisch ist. Bevor die Daten in die SelenoDB integriert werden konnten, mussten die SNP-Dateien pro Spezies per Gen geteilt werden.

Als letzter Schritt wurde eine Tabelle erstellt, die die Anzahl von SNPs pro Gen visualisiert.

## **References**

1. Wang J, Schierupc M H, Extensive X-linked adaptive evolution in central chimpanzees, PNAS, 2011
2. Volff, J-N, Gene and Protein Evolution
3. Tulkem R, Behavior, Ecology and Conservation, Springer, 2008
4. The Chimpanzee Sequencing and Analysis Consortium(Tarjei S et al), Initial sequence of the chimpanzee genome and comparison with the human genome, Nature, 2005
5. Stadtman T C, Selenocysteine. Annu Rev Biochem, 1996
6. Springer M S, Meredith RW, Gatesy J, Emerling CA, Park J,
7. Macroevolutionary Dynamics and Historical Biogeography of Primate Diversification Inferred from a Species Supermatrix. PLoS ONE, 2011
8. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K, db SNP: The NCBI database of genetic variation, Nucleic Acids Res, 2001
9. Robinson J, Thorvaldsdóttir H, Winckler W, Guttman M, Lander E, Getz G ,Mesirov J, Integrative genomics viewer, Nature Biotechnology, 2011
10. Richard A, Primates in Nature, W.H.Freeman and Company New York,1985
11. Quinlan AR, Hall IM, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics, 2010
12. Quinlan A, Hall I, BEDTools: a flexible suite of utilities for comparing genomic features, Oxfordjournal,, 2010
13. Prüfer K, Munch K, Hellmann I, Akagi K, Miller J R, Walenz B, Koren S, Sutton G, Kodira C, Winer R, Knight J R, Mullikin J C, Meader S J, Ponting C P, Lunter G, Higashino S, Hobolth A, Dutheil J, Karakoç E, Alkan C, Sajjadian S, Catacchio C R, Ventura M, Marques-Bonet T, Eichler EE , The bonobo genome compared with the chimpanzee and human genomes, Nature, 2012
14. Pitts MW, Raman AV, Hashimoto AC, Todorovic C, Nichols RA, Berry MJ Deletion of selenoprotein P results in impaired function of parvalbumin interneurons and alterations in fear learning and sensorimotor gating, Epub, 2012.
15. Pitt J, Ferré-D'Amaré A, Rapid Construction of Empirical RNA Fitness Landscapes, Science, 2010
16. Pattnaik S, Vaidyanathan S, Pooja D G, Deepak S, Panda B, Customization of the Exome Data Analysis Pipeline using a combinatorial Approach, PLOS ONE, 2012

17. Parla J S, Iossifov I, Grabill I, Spector M S, Kramer M, McCombie W R, A comparative analysis of exome capture, *Genome Biology* 2011
18. Ng S B, Turner E H, Robertson P D, Flygare S D, Bigham A W, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler E E, Bamshad M, Nickerson D A, Shendure J, Targeted capture and massively parallel sequencing of 12 human exomes, *Nature*, 2009
19. Neumann R, SNP (Single Nucleotide Polymorphism), *Laborjournal-Edition* 10, 2000, (Last change: 20.10.2004)
20. Nei M, *Molecular Evolutionary Genetics*, Columbia University Press, 1987
21. Mielke J, *Human Biological Variation*, Oxford University Press, 2006
22. Mc Kenna A, Hanna M, The Genome Analysis Toolkit: A Mapreduce framework for analyzing next-generation-Sequencing data, *Nature*, 09.2010
23. Maurano M, Systematic Localization of Common Disease Associated Variation in Regulator DNA, *Science*, 2012
24. Li D, Zheng D H, Jiang D T, Jiang D H, Jin X, Munch K, Hobolth A, Siegmund HR, Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, 2009
25. Kwok P Y, Chen X, *Detection of Single Nucleotide Polymorphisms*, Caister Academic Press, 2003
26. Kryukov G V, Castellano S, Novoselov S V, Lobanov A V, Zehrab O, Guigo R, Gladyshev V N, Characterization of Mammalian Selenoproteomes, *Science*, 2003
27. Jennifer S Parla<sup>†</sup>, Ivan Iossifov<sup>†</sup>, Ian Grabill, Mona S Spector, Melissa Kramer and W Richard McCombie, A comparative analysis of exome capture, *Genome Biology*, 2011
28. Jackson M, *Human Genome Evolution*, BIOS Scientific Publishers, 1996
29. Hviil C, Qian C, Bataillon T, Li Y, Mailund T, Sallé B, Carlsén F, Hartl D, Clark A, *Principles of Population Genetics* 4<sup>th</sup> Edition, Sinauer Associates, Inc Publishers, Sunderland Massachusetts, 2006
30. Gonzalez-Flores J N, Gupta N, Demong L W, Copeland P R, The Selenocysteine-specific Elongation Factor Contains a Novel and Multi-functional Domain, *J Biol Chem*, 2012
31. Goecks J, Nekrutenko A, Taylor J, The Galaxy Team, *Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences*. *Genome Biol*, 2010



- 
32. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. "Galaxy: a platform for interactive large-scale genome analysis." *Genome Research*, 2005
  33. Freenab S, Herron J, *Evolutionary Analysis* 2<sup>nd</sup> Edition, Prentice Hall College Div, 2000
  34. Eckstein F, Lilley, D, *Nucleic Acids and Molecular Biology*, Springer, 1998
  35. Durbin R, Haussler D, The default version for GFF files is now Version 2, published online, 2000, <http://www.sanger.ac.uk/resources/software/gff/spec.html> (accessible 06.12.2012)
  36. Danecek P, Auton A, Abecasis G, Albers C A, Banks E, DePristo M, Handsaker R, Lunter G, Marth G, Sherry S T, McVean G, Durbin R, The variant call format and VCFtools, *Bioinformatics*, 2011
  37. Castellano S, Andrés M A, Bosch E, Bayes M, Guigó R, Clark A G, Low Exchangeability of Selenocysteine, the 21st Amino Acid, in Vertebrate Proteins, *Mol Biol Evol*, 2009
  38. Brown, T A, *Genomes3*, Garland Science: New York, 2007
  39. Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J, *Galaxy: a web-based genome analysis tool for experimentalists*, *Current Protocols in Molecular Biology*, 2010
  40. Barnes M, Gray Wiler I, *Bioinformatics for Geneticists* 1<sup>st</sup> Edition, *Bioinformatics for Geneticists*, 2003
  41. Anapol F, *Shaping Primate Evolution*, Cambridge University Press, 2004
  42. Altshuler D, Daly M J, Lander E S, *Genetic Mapping in Human Disease*, *Science*, 2008:
  43. Albert T J, Molla M N, Muzny D M, Nazareth L, Wheeler D, Song X, Richmond T M, Middle C M, Rodesch M J, Packard C J, Weinstock G M, Gibbs R A, Direct selection of human genomic loci by microarray hybridization, *Nat Methods*, 2007

## **Appendix**

List of all genes whose regions were used to call for SNPs

G6PD	SEPHS1
SEPN1	MSRB2
TRNAU1AP	PSTK
GPX7	CELF1
LRP8	C11orf31
DIO1	MSRB3
SEP 15	TXNRD1
SELENBP1	GPX2
EPT1	DIO2
XPO1	DIO3
LRP2	SELS
SCLY	SEPX1
GPX1	SEPHS2
SELK	C17orf37
TXNRD3	EIF4A3
EEFSEC	GPX4
SELT	ELAVL1
SEPSECS	SARS2
SEPP1	SELV
GPX8	SEPW1
GPX3	TXNRD2
GPX6	SELM
GPX5	SELO
TTPA	17.tRNA25-SeCTCA
RPL30	19.tRNA8-SeC(e)TCA
SECISBP2	22.tRNA1-SeC(e)TCA

**Statement of authorship (Selbstständigkeitserklärung)**

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Literatur und Hilfsmittel angefertigt habe.

Stellen, die wörtlich oder sinngemäß aus Quellen entnommen wurden, sind als solche kenntlich gemacht.

Diese Arbeit wurde in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt.

Leipzig, den 13.12.2012

Ron Hübler